(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2001－13982
（P2001－13982A）
(43)公開日　平成13年１月19日(2001.1.19)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ１０Ｌ　13/06		Ｇ１０Ｌ　5/04		Ｆ　５Ｄ０４５
13/08		3/00		Ｈ　９Ａ００１

審査請求　未請求　請求項の数６　ＯＬ　（全 23 頁）

(21)出願番号　特願平11－219216

(22)出願日　平成11年８月２日(1999.8.2)

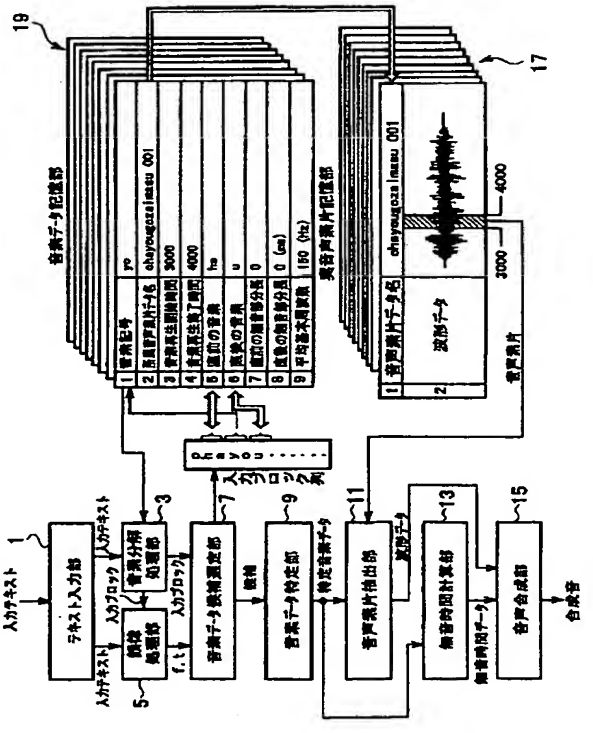
(31)優先権主張番号　特願平11－122718
(32)優先日　平成11年４月28日(1999.4.28)
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000004329
日本ビクター株式会社
神奈川県横浜市神奈川区守屋町３丁目12番地
(72)発明者　和田　祐司
神奈川県横浜市神奈川区守屋町３丁目12番地　日本ビクター株式会社内
(74)代理人　100083806
弁理士　三好　秀和　（外９名）

Ｆターム(参考)　5D045 AA08 AA09 AA11
9A001 HH18 HZ30

(54)【発明の名称】　音声合成装置

(57)【要約】
【課題】　音声素片の接続部分のノイズを低減するとともに、録音された実音声の特徴を備えた高品質の合成音出力が可能な音声合成装置を提供する。
【解決手段】　音素データ記憶部１９の各保存領域には、「音素記号」、「所属音声素片データ名」、「直前の音素」、「直後の音素」が予め登録され、実音声素片記憶部１７には、実音声を記憶した波形データが音声素片データ名とともに蓄積して記憶される。音素データ特定部９は、入力ブロック列のなかの互いに隣接する「入力ブロック」の組と、音素データ記憶部１９の「音素記号」および「直前の音素」（あるいは、「直後の音素」）とを照合することにより、各入力ブロックに対して隣接音声環境が最適な音素データを特定し、音声素片抽出部１１は特定音素データの「所属音声素片データ名」を基に実音声素片記憶部１７から音声素片を抽出し、音声合成部１５は抽出された音声素片を順次接続し合成音を出力する。



1

【特許請求の範囲】

【請求項１】　音声波形が予め記憶された波形記憶手段と、

該波形記憶手段の音声波形に含まれる第１の音声素片に対して付与された第１のラベルと、当該第１の音声素片に隣接する音声素片に対して付与された第２のラベルと、当該第１の音声素片の記憶場所情報とを含む音素データが予め記憶される音素データ記憶手段と、

音声合成指令としての入力音素記号列と該入力音素記号列に隣接する隣接入力音素記号列とからなる組と、前記音素データに含まれる第１のラベルおよび第２のラベルの組とを照合して、当該入力音素記号列に対する音素データを特定する音素データ特定手段と、

前記波形記憶手段のなかの前記特定された音素データの示す記憶場所に記憶された音声素片を抽出する音声素片抽出手段と、

該音声素片抽出手段で抽出された音声素片を所定の順序で接続し出力する音声合成手段とを有する音声合成装置。

【請求項２】　請求項１記載の音声合成装置において、前記音素データには、前記第１の音声素片と該第１の音声素片に隣接する音声素片との間の状態に係る時間情報が付与され、該時間情報に基づいて音声素片の接続部分の状態を設定する時間設定手段を有することを特徴とする音声合成装置。

【請求項３】　請求項２記載の音声合成装置において、前記時間設定手段は、前記音素データに、前記第１の音声素片と該第１の音声素片の前後に隣接する各音声素片との間の無音部分の時間情報がそれぞれ付与されているとき、前記第１の音声素片の後の無音部分の時間情報と、当該第１の音声素片の後に接続される音声素片の前の無音部分の時間情報の少なくとも一方に対し重み付けを行って、当該各音声素片を接続する部分に設ける無音時間を算出することを特徴とする音声合成装置。

【請求項４】　請求項１ないし請求項３のいずれかに記載の音声合成装置において、

前記入力音素記号列に対して特定された音素データが、当該入力音素記号列の前又は後ろに隣接する入力音素記号列に対して特定されることを規制する規制手段を有することを特徴とする音声合成装置。

【請求項５】　請求項１ないし請求項４のいずれかに記載の音声合成装置において、

前記音声合成手段により出力される音声素片の出力タイミングに係る情報を生成するタイミング情報生成手段を具備することを特徴とする音声合成装置。

【請求項６】　請求項５記載の音声合成装置において、前記タイミング情報生成手段は、前記音声合成手段により出力される音声素片に当該音声素片よりも短い音声が含まれるときには、当該音声との同期をとるための情報を前記出力タイミングに係る情報に含ませることを特徴とする音声合成装置。

2

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、音声素片の接続部分のノイズを低減するとともに、録音された音声の特徴を備えた高品質の合成音出力が可能な音声合成装置および合成音に同期させて、顔画像における口の形状を変化させることのできる音声合成装置に関する。

【０００２】

【従来の技術】近年、人の実音声を録音して保存したものから合成音を生成する録音合成の技術が、音声の特徴の再現性の点から注目されている。

【０００３】従来の録音合成を利用した音声合成装置にあっては、予め実音声を録音し、例えば、「あ」や「い」等の５０音を構成する音声等の所定の音声単位ごとに区切って多数の音声素片としたものが用意され、この各音声素片は、音声素片の種類を表す音素記号、音声の高低を決める基本周波数および再生時間長が付与されてデータベースとして蓄積保存される。

【０００４】テキストから音声を合成する際には、入力された日本語テキストは、先ず単語辞書を基にした音素分解処理により音素記号に分解され、その得られた各音素記号に対して、所定のアクセントパターン表および再生時間長規則を基にした韻律処理により基本周波数と再生時間長とが割り当てられる。そして、音素記号、基本周波数および再生時間長と、上述のデータベースのなかの音素記号、基本周波数および再生時間長とが照合され、音素分解処理により得られた音素記号に対して音声素片が選択される。そして、選択された音声素片が入力されたテキストの順番で接続されて合成音声が生成される。

【０００５】また、音声合成の技術が注目されるなかにあって、その一方では、音声解析やＣＧ（Computer Graphics）の技術の加速的な進歩により、入力された音声波形を解析しながら、実際に人が音声をしゃべっているように口の形状が変化する顔画像を生成する画像処理技術が実現段階に入っている。このような技術は、例えば、事件の現場から電話により伝えられた言葉を、ＣＧで制作されたアナウンサにしゃべらせるというような用途などに利用される。

【０００６】このような用途において、顔画像における口の形状をリアルタイムに変化させるためのパラメータを生成するためには、コンピュータシステムなどを利用して、入力された音声波形の解析処理が行われる。

【０００７】

【発明が解決しようとする課題】しかしながら、上記従来の音声合成装置では、合成音声の生成のための音声素片を記憶するにあたっては、実音声において隣接する音声素片に関係なく互いに分離され、一方、音声素片を合成するにあたっては、入力された音素記号等と、データ

3

ベースの各音声素片に対応づけられた音素記号等とを照合することにより音声素片を選択し、その選択された音声素片を単純に接続する方法が採られていたため、生成される合成音は、音声素片の接続部分に生ずるノイズが高く、録音された実音声の特徴が損われてしまう場合があった。

【０００８】また、上記のように、顔画像における口の形状を音声に同期して変化させるためには、処理能力の高いコンピュータシステムによる複雑で膨大な音声解析処理が必要であることから、このような音声解析処理を行わずに、顔画像における口の形状を音声と同期したタイミングで変化させたいとの要望が多くなっている。

【０００９】そこで本発明は、上記従来の問題点に鑑みてなされたものであり、その目的とするところは、音声素片の接続部分のノイズを低減するとともに、録音された実音声の特徴を備えた高品質の合成音出力が可能な音声合成装置を提供することにある。

【００１０】また、本発明の他の目的とするところは、合成音に同期させて、顔画像における口の形状を変化させることのできる音声合成装置を提供することにある。

【００１１】

【課題を解決するための手段】本発明の請求項１に係る音声合成装置は、音声波形が予め記憶された波形記憶手段と、該波形記憶手段の音声波形に含まれる第１の音声素片に対して付与された第１のラベルと、当該第１の音声素片に隣接する音声素片に対して付与された第２のラベルと、当該第１の音声素片の記憶場所情報とを含む音素データが予め記憶される音素データ記憶手段と、音声合成指令としての入力音素記号列と該入力音素記号列に隣接する隣接入力音素記号列とからなる組と、前記音素データに含まれる第１のラベルおよび第２のラベルの組とを照合して、当該入力音素記号列に対する音素データを特定する音素データ特定手段と、前記波形記憶手段のなかの前記特定された音素データの示す記憶場所に記憶された音声素片を抽出する音声素片抽出手段と、該音声素片抽出手段で抽出された音声素片を所定の順序で接続し出力する音声合成手段とを有する。

【００１２】本発明の請求項１に係る音声合成装置では、音声波形が予め記憶された波形記憶手段が設けられ、さらに、この波形記憶手段の音声波形に含まれる第１の音声素片に対して付与された第１のラベルと、第１の音声素片に隣接する音声素片に対して付与された第２のラベルと、第１の音声素片の記憶場所情報とを含む音素データが予め記憶される音素データ記憶手段が設けられ、音声合成指令として入力音素記号列が入力されると、音素データ特定手段により、その入力音素記号列と隣接入力音素記号列とからなる組と、音素データに含まれる第１のラベルおよび第２のラベルの組とが照合され、入力音素記号列に対する音素データが特定され、その後、音声素片抽出手段により、波形記憶手段のなかの

4

特定された音素データの示す記憶場所から、逐次音声素片が抽出され、音声合成手段により、適切な合成音となるような順序で、これら音声素片が接続され、合成音が出力される。

【００１３】また、本発明の請求項２に係る音声合成装置は、請求項１記載の音声合成装置において、前記音素データには、前記第１の音声素片と該第１の音声素片に隣接する音声素片との間の状態に係る時間情報が付与され、該時間情報に基づいて音声素片の接続部分の状態を設定する時間設定手段を有することを特徴とする。

【００１４】本発明の請求項２に係る音声合成装置では、音素データには、第１の音声素片と、該第１の音声素片との間の状態に係る時間情報が予め付与され、時間設定手段が時間情報に基づいて、音声素片の接続部分の状態を設定して、音声上の人の息継ぎ等の特徴が再現可能となる。

【００１５】また、本発明の請求項３に係る音声合成装置は、請求項２記載の音声合成装置において、前記時間設定手段は、前記音素データに、前記第１の音声素片と該第１の音声素片の前後に隣接する各音声素片との間の無音部分の時間情報がそれぞれ付与されているとき、前記第１の音声素片の後の無音部分の時間情報と、当該第１の音声素片の後に接続される音声素片の前の無音部分の時間情報の少なくとも一方に対し重み付けを行って、当該各音声素片を接続する部分に設ける無音時間を算出することを特徴とする。

【００１６】本発明の請求項３に係る音声合成装置では、音素データには、予め第１の音声素片と該第１の音声素片の前後に隣接する各音声素片との間の無音部分の時間情報がそれぞれ付与され、時間設定手段により、第１の音声素片の後の無音部分の時間情報と、当該第１の音声素片の後に接続される音声素片の前の無音部分の時間情報の少なくとも一方に対して重み付けが行われて、各音声素片を接続する部分に設ける無音時間が算出され、どちらか一方の音声素片の特徴に偏った無音時間が設定されることが回避され、たとえ実音声上で隣接していない音声素片どうしを接続する際にあっても、自然な合成音が得られる。

【００１７】また、本発明の請求項４に係る音声合成装置は、請求項１ないし請求項３のいずれかに記載の音声合成装置において、前記入力音素記号列に対して特定された音素データが、当該入力音素記号列の前又は後ろに隣接する入力音素記号列に対して特定されることを規制する規制手段を有することを特徴とする。

【００１８】本発明の請求項４に係る音声合成装置では、入力音素記号列に対して特定された音素データは、当該入力音素記号列の前又は後ろに隣接する入力音素記号列に対して特定されずに、実音声上では連続することが極めて希な同一の音声素片が特定されて合成音が不自然となる不都合が回避される。

【００１９】また、本発明の請求項５に係る音声合成装置は、請求項１ないし請求項４のいずれかに記載の音声合成装置において、前記音声合成手段により出力される音声素片の出力タイミングに係る情報を生成するタイミング情報生成手段を具備することを特徴とする。

【００２０】本発明の請求項５に係る音声合成装置では、音声合成手段により出力される音声素片の出力タイミングに係る情報が、タイミング情報生成手段によって生成され、この情報によって、顔画像における口の形状を変化させるタイミングを決定することができるようにしている。

【００２１】また、本発明の請求項６に係る音声合成装置は、請求項５記載の音声合成装置において、前記タイミング情報生成手段は、前記音声合成手段により出力される音声素片に当該音声素片よりも短い音声が含まれるときには、当該音声との同期をとるための情報を前記出力タイミングに係る情報に含ませることを特徴とする。

【００２２】本発明の請求項６に係る音声合成装置では、音声合成手段により出力される音声素片に、例えば母音や子音などの音素や母音および子音からなる音節などの、当該音声素片よりも短い音声が含まれるときには、タイミング情報生成手段により、当該音声との同期をとるための情報を出力タイミングに係る情報に含ませて、顔画像における口の形状を音声素片よりも短い音声に同期させて変化させられるようにしている。

【００２３】

【発明の実施の形態】以下、本発明に係る音声合成装置の実施の形態を図１ないし図１１を参照して説明する。尚、以下の説明では、日本語テキストからの音声合成を例にするが、音素分解処理部および韻律処理部を変えれば、他の言語で入力されたテキストからの音声合成が可能である。

【００２４】図１は、本発明に係る音声合成装置の第１の実施の形態を示す図である。

【００２５】本実施の形態の音声合成装置には、キーボードや、予め用意されたテキストデータを保存するハードディスク装置等の外部記憶装置が接続される。そして、この音声合成装置には、所定の処理を実行する演算部、および処理命令の記憶やデータの一時保存が可能な主記憶部とが設けられており、この主記憶部や、外部記憶装置に記憶された命令を逐次演算部に読み込ませ、実行させることにより以下の音声合成のための処理が行われる。また、以下に説明する各部間のデータの転送については、主記憶部に設けられた所定の保存領域、或いは命令によって逐次設定される作業領域を介して行われる。

【００２６】尚、必要に応じてマウス、ディスプレイ装置等の入出力装置を接続してもよい。また、後述する実音声波形記憶部１７と音素データ記憶部とを、複数の端末との通信が可能なコンピュータ等に配備して、ネットワーク型の音声合成装置を構成することも可能である。

【００２７】本図に示すテキスト入力部１は、キーボードや外部記憶装置等から、入力バッファを介して、日本語テキストデータを入力し、そのデータを音素分解処理部３、および韻律処理部５へと出力する。音素分解処理部３は、テキスト入力部１から転送された日本語テキストデータに言語処理を行うとともに、音声素片（この音声素片には、単音素および、単音素が複合された音素列が含まれる）の種類を表し、予め後述する音素データ記憶部１９に登録されている音素記号あるいは音素記号列に変換する。

【００２８】以下の説明では、入力されたテキストデータを「入力テキスト」といい、この変換により得られた各音素記号列（単一の音素記号を含む）を「入力ブロック」ないしは「入力音素記号列」という。例えば、入力テキスト「おはようございます」は、この時点で８個の入力ブロック／ｏ／，／ｈａ／，／ｙｏ／，／ｕ／，／ｇｏ／，／ｚａｉ／，／ｍａ／，／ｓｕ／からなる入力ブロック列に変換される。

【００２９】そして、音素分解処理部３は、各入力ブロックを韻律処理部５および音素データ候補選定部７へと送出する。韻律処理部５は、所定のアクセントパターン表や再生時間長規則を基にした韻律処理を実行することにより、音素分解処理部３から入力された各入力ブロックに対して、テキスト入力部１から入力された入力テキストを解析しながら、平均基本周波数ｆと再生時間長ｔを決定する。そして、韻律処理部５は、各入力ブロックに平均基本周波数ｆと再生時間長ｔとを対応づけて、音素データ候補選定部７へと送出する。

【００３０】音素データ候補選定部７は、入力された入力ブロック、平均基本周波数ｆおよび再生時間長ｔと、後述する音素データ記憶部１９に記憶されたデータベースに含まれるデータとを照合して、各入力ブロックに対する音素データの候補を選定する。そして、候補を表すフラグを音素データに対応づけ、後述する音素データ特定部９での処理において、候補とそうでないものの区別を可能とする。

【００３１】音素データ特定部９は、前記候補を表すフラグに基づいて、各入力ブロックについての候補のなかから、合成音生成のために最適な１つの音素データを最終的に特定するまでの処理を行う。これ以降、特定された音素データを特定音素データという。この処理により、各入力ブロックに対する、合成音生成のための音声素片が抽出可能となる。そして、特定音素データを後述する音声素片抽出部１１および無音時間計算部１３へと出力する。

【００３２】音声素片抽出部１１は、特定音素データの内容を解析して、合成音を生成するための音声素片を実音声波形記憶部１７から抽出する。そして音声素片を音声合成部１５へと出力する。

【００３３】無音時間計算部１３は、音素データ特定部９から入力された特定音素データの内容を解析し、各入力ブロックに対する音声素片を接続する際の、接続部分に設ける無音時間を計算する。そして、無音時間計算部１３は、無音時間を示す無音時間データを音声合成部１５に送出する。

【００３４】音声合成部１５は、無音時間計算部１３から入力された無音時間データに基づいて、無音時間を音声素片波形データの接続部に設定して、音声素片抽出部１１から入力した各波形データと各無音時間を順次接続することで、１つの合成音データを生成し、これを増幅して出力する。

【００３５】実音声波形記憶部１７は、図１に示すように、合成音生成の際に音声素片抽出部１１により抽出される音声素片を含む波形データがデータベースとして蓄積記憶された部分である。これら波形データは、例えば、小説等の所定の日本語長文を人間に読ませ、これを録音して適当な長さに分割したものを符号化したものであり、様々な録音時間の波形データが含まれている。また、同じ音声からのデータであっても、録音時の声の状態や、読まれた文章の内容等により、その波形は異なったものになる。

【００３６】この実音声波形記憶部１７は、各データ毎に、２つの保存領域を有し、保存領域１「音声素片データ名」には、検索用の名前が保存されている。図１の例では、音声「おはようございます」に相当する名前「ｏｈａｙｏｕｇｏｚａｉｍａｓｕ００１」が保存されている。また、保存領域２「波形データ」には、音声「おはようございます」に相当する波形データが保存されている。

【００３７】そして、本図に示す音素データ記憶部１９は、保存領域１から保存領域９までの９つの保存領域を有する音素データのデータベースである。この音素データ記憶部１９は、音声素片に１対１に対応している（以下、この音声素片を、各保存領域の説明において「当該音声素片」という）。そして、音素分解処理の際には、入力テキストに対して入力ブロックを決定（割り当て）するために、音素分解処理部３によって参照される部分であり、また、音素データ特定部９によって、最適な音声素片を特定するために照合される部分でもある。

【００３８】その保存領域１「音素記号」には、入力ブロックと対比される音素記号（または音素記号列）が登録保存される。本例では、音声「よ」に相当する音素記号／ｙｏ／が保存されている。即ち、この音素記号または音素記号列が第１のラベルとしての機能を果す。

【００３９】保存領域２「所属音声素片データ名」には、当該音声素片が、実音声波形記憶部１７に生成されたデータベースの中の、どのデータに所属するかを示す名前が保存される。本図の例では、実音声「おはようございます」の波形データを保存した「ｏｈａｙｏｕｇｏｚａｉｍａｓｕ００１」というデータの中に、音素記号／ｙｏ／に相当する音声素片が含まれていることを示している。

【００４０】保存領域３「音素再生開始時間」には、当該音声素片が、「所属音声素片データ名」の名前の示す再生音声（以下、各保存領域の説明において「当該再生音声」という）の、どの時点から開始される音声に相当するものかを示すデータが保存される。このデータとしては、トータルの再生時間に対する相対的な開始時刻を示す値等を使用すればよい。

【００４１】一方保存領域４「音素再生終了時間」には、当該音声素片が、当該再生音声の、どの時点までの音声に相当するものかを示すデータが保存される。この「音素再生終了時間」にも「音素再生開始時間」と同様に、トータルの再生時間に対する相対値等を使用すればよい。従って、「所属音声素片データ名」、「音素再生開始時間」および「音素再生終了時間」により当該音声素片の記憶場所が特定可能となっている。

【００４２】そして、保存領域５「直前の音素」には、当該再生音声のなかで、当該音声素片の直前の音声素片を示す音素記号（または音素記号列）が登録保存され、一方保存領域６「直後の音素」には、当該音声素片の直後に再生される音声素片を示す音素記号（または音素記号列）が登録保存される。即ち、これら保存領域「直前の音素」または「直後の音素」に登録された音素記号（または音素記号列）が第２のラベルとして機能する。

【００４３】保存領域７「直前の無音部分長」には、当該音声素片とその直前の音声素片とのとの間の無音部分の長さを示すデータが保存され、一方保存領域８「直後の無音部分長」には、当該音声素片とその直後の音声素片とのとの間の無音部分の長さを示すデータが保存される。これら「直前の無音部分長」および「直後の無音部分長」には、実際の時間情報を保存しても良いし、また、当該音声素片の再生時間等に対する相対値を保存してもよい。尚、これら無音部分は、ポーズともいわれる。そして、音素データ記憶部１９の保存領域９「平均基本周波数」には、当該音声素片の平均基本周波数を示す値が保存される。

【００４４】次に、図２に示すフローチャートを参照して、本発明に係る音声合成装置の動作を実例をまじえて説明する。本図に示す処理は、入力テキストが音声合成装置の入力バッファ等に蓄えられている状態から、音声合成の指令等によって開始するものである。尚、以下の実施の形態の説明において、データベースの保存領域名を鍵かっこで囲んだ表現を使用する場合は、データベースの保存領域に保存されたデータそのものを示すものとする。

【００４５】図２においては、先ず、ステップＳ１で、例えば「おはようございます」等の入力テキストが、入力バッファ等からテキスト入力部１へと入力される。入

力された文字列は、ステップＳ３において、音素分解処理部３へと出力され、音素分解処理部３では、音素データ記憶部１９に保存された「音声」を参照することにより、入力テキストを複数の入力ブロックに変換する（この変換処理を音素分解という）。

【００４６】上記例示した入力テキスト「おはようございます」は、この時点で８個の入力ブロック／ｏ／，／ｈａ／，／ｙｏ／，／ｕ／，／ｇｏ／，／ｚａｉ／，／ｍａ／，／ｓｕ／からなる入力ブロック列に変換される。尚、便宜上、得られた入力ブロックの数をｎ個とし、各入力ブロックをそれぞれ、入力テキストの順に合わせて、第１入力ブロックｃ１、第２入力ブロックｃ２，…，第ｎ入力ブロックｃｎというものとする。

【００４７】韻律処理部５は、次のステップＳ５において、各入力ブロックに対して、所定のアクセントパターン表や再生時間長規則に基づいて、音声再生時の音程を決定する平均基本周波数ｆと音声の継続時間を決定する再生時間長ｔとを割当てる。ここで、ｉ番目の入力ブロックに割当てられた平均基本周波数、音素再生長をそれぞれｆｉ、ｔｉというものとする。尚、入力ブロックが単一の音素記号の複合されたものである場合は、その音素記号列の各音素記号に割当てられた基本周波数を平均化したものが平均基本周波数ｆとなる。

【００４８】そして、韻律処理部５は、平均基本周波数ｆと再生時間長ｔとを入力ブロックに対応づけて、主記憶部の所定の保存領域に保存する。尚、入力ブロック、平均基本周波数および再生時間長を総称して入力データという。

【００４９】そして、図２のステップＳ７においては、音素データ候補選定部７は、第１入力データ、即ち、第１入力ブロック、平均基本周波数ｆ１および再生時間長ｔ１を読み込み、所定の作業領域に保存する。ステップＳ９においては、音素データ候補選定部７は、各入力データを音素データ記憶部１９に記憶されたデータに対比させて、音素データの候補を選定する。

【００５０】ここで、図３に示すフローチャートを参照して、音素データ候補選定部７にて各入力ブロックについて実行される、音素データの候補選定処理について説明する。この候補選定処理は、最終的に、入力ブロックに対して１つの音素データを特定するための前段階として、実行されるものであり、先ず、「音声記号」が入力ブロックに一致し、しかも「平均基本周波数」および音声素片の再生時間が、入力ブロックに割り当てられたものに対して、所定の誤差範囲に含まれる音素データが候補として選定される。選定にあたっては、候補を表すフラグを音素データに対応づけ、後の処理のために、このフラグをオン／オフいずれかの状態に設定する。

【００５１】図３におけるステップＳ４１では、音素データ候補選定部７は、入力ブロックｃｉを検索パラメータとして音素データ記憶部１９に記憶された全音素データの保存領域１を検索し、入力ブロックと「音声記号」が一致する音素データを候補として選定する。音素データ記憶部１９には、通常「音素記号」が同一である多数の音素データが保存されているため、通常は、この時点での候補数は比較的多い。

【００５２】音素データ候補選定部７は、前述のフラグの状態をもとに、すべての候補について、順次に以下の処理を行う。

【００５３】先ずステップＳ４３において、第１の候補（音素データ）を読み込んで主記憶部の所定の作業領域に保存する。そして、ステップＳ４５においては、当該入力ブロック（ｃｉ）に割当てられた平均基本周波数ｆｉと、当該音素データの「平均基本周波数」との差Δｆ（偏差）が所定の周波数Ｆ以上であるか否かを判定する。このステップＳ４５で、この偏差Δｆが値Ｆ以上であると判定される（ＹＥＳ）と判定されると、ステップＳ４７において、この音素データを候補から除外する処理（以下、候補除外処理という）が行われる。即ち、「平均基本周波数」が所定の誤差範囲にない音素データは、ここで候補から除外される。

【００５４】一方、ステップＳ４５で、偏差Δｆが周波数Ｆより小さい（ＮＯ）と判定されると、続くステップＳ４９においては、当該対象となっている候補の「音素再生開始時間」と「音素再生終了時間」との時間差、即ち音声素片の再生時間長と、入力ブロックｃｉに割当てられた再生時間長ｔｉとの偏差Δｔが所定の時間値Ｔ以上であるか否かを判定する。このステップＳ４９において、偏差Δｔが値Ｔ以上である（ＹＥＳ）と判定されると、ステップＳ４７において、候補除外処理が行われる。即ち、再生時間長が所定の誤差範囲にない音素データは、ここで候補から除外される。

【００５５】一方、ステップＳ４９において、偏差Δｔが値Ｔより小さい（ＮＯ）と判定された場合は、次にステップＳ５１にて、その候補（音素データ）が当該入力ブロックｃｉの直前の入力ブロックｃｉ－１について特定された音素データであるか否かが判定され、直前の入力ブロックｃｉ－１について特定された音素データである（ＹＥＳ）と判定されたときは、制御がステップＳ４７に移行して、その候補について候補除外処理が行われる。即ち、実音声では、同一音が連続することが極めて希であるため、音声合成装置がこのような誤った選定を行い、合成音が不自然となるのを回避する効果がある。

【００５６】ステップＳ５１にて、直前の入力ブロックｃｉ－１について特定された音素データでない（ＮＯ）と判定されたときは、その音素データは候補として残されることとなる。そして、ステップＳ５１にてＮＯと判定された後、あるいは、ステップＳ４７における候補除外処理の終了後は、音素データ候補選定部７は、ステップＳ５３において、当該候補が最後の候補か否かを判定する。ここで、最後の候補でない（ＮＯ）と判定される

と、音素データ候補選定部７は、ステップＳ５５に制御を移行させ、次の候補（音素データ）の内容を読み込む。そして、音素データ候補選定部７は、ステップＳ４５からステップＳ５３までの一連の処理を、ステップＳ５３において、最後の候補である（ＹＥＳ）と判定されるまで順次実行する。

【００５７】音素データ候補選定部７が上記処理を行うことにより、後述する波形の抽出時において、同一の音素記号で、しかも音声の平均基本周波数および再生時間長の近い音声素片の抽出が可能となる。

【００５８】そして、図２に示すしたステップＳ１１においては、例えば、図１に示したような１つの入力ブロック「ｙｏ」に対して、ある程度絞り込まれた候補のなかから最終的な１つの音素データを特定する処理を音素データ特定部９が行う。特にこの実施の形態では、入力ブロックの前後の関係を、波形データ記録時の音声素片の前後関係に対比させることにより、隣接音声環境の最適な音声素片を特定することを可能としている。

【００５９】図４に示すフローチャートは、第１入力ブロックに対して、音素データ特定部９が実行する処理を説明するためのものであり、本図のステップＳ６１においては、第２入力ブロックを主記憶部の所定作業領域に読み込む。そして、現在残っている候補のなかの最初の候補（音素データ）につき、その「直後の音素」を読み込む（ステップＳ６３）。そして、ステップＳ６１にて読み込んだ入力ブロックｃ２とステップＳ６３にて読み込んだ「直後の音素」とが一致するか否かを、ステップＳ６５にて判定し、ここで一致しない（ＮＯ）と判定された場合は、音素データ特定部９は、ステップＳ６７において候補除外処理を行う。

【００６０】一方、ステップＳ６５にて、一致する（ＹＥＳ）と判定されたときは、候補除外処理は行われない。ステップＳ６５にて、一致する（ＹＥＳ）と判定された後、あるいは、ステップＳ６７における候補除外処理の終了後は、音素データ特定部９は、ステップＳ６９において、当該候補が最後の候補か否かを判定する。ここで、最後の候補でない（ＮＯ）と判定されると、ステップＳ７１に進んで、次の候補の「直後の音素」を読み込み、そして、音素データ特定部９は、ステップＳ６５からステップＳ６９までの一連の処理を、ステップＳ６９において、最後の候補である（ＹＥＳ）であると判定されるまで順次実行する。

【００６１】このようにして、第１入力ブロックについては、当該候補の「直後の音素」が第２入力ブロックと一致するものだけが、候補として残されることとなる。

【００６２】図５に示すフローチャートは、第１入力ブロックおよび最後の入力ブロックを除く、任意の第ｉ入力ブロックに対して、音素データ特定部９が実行する処理を説明するためのものである。

【００６３】音素データ特定部９は、ステップＳ８１において、当該入力ブロックｃｉの直前の入力ブロックおよび直後の入力ブロック、即ち、入力ブロックｃｉ－１および入力ブロックｃｉ＋１を主記憶部に読み込み、その後、ステップＳ８３において、最初の候補の「直前の音素」および「直後の音素」を主記憶部に読み込む。そして、ステップＳ８５においては、当該候補（音素データ）が以下のどの条件を満たすかにより候補の分類を行う。

【００６４】先ず、当該候補の「直前の音素」と入力ブロックｃｉ－１とが一致し、かつ、「直後の音素」と入力ブロックｃｉ＋１が一致する候補に対しては、優先して候補に残される可能性（優先度）が一番高いことを示すデータＰ１が対応づけられる。図１に示す例では、入力ブロック「ｙｏ」に対して、「直前の音素」が「ｈａ」であり、かつ、「直後の音素」が「ｕ」であり、この条件を満たすため、この音素データには、データＰ１が対応づけられる。

【００６５】そして、「直前の音素」と入力ブロックｃｉ－１が一致し、かつ、「直後の音素」と入力ブロックｃｉ＋１が一致しない候補に対しては、データＰ１よりも優先度が低いことを示すデータＰ２が対応づけられる。

【００６６】そして、さらに「直前の音素」と入力ブロックｃｉ－１とが一致せず、かつ、「直後の音素」と入力ブロックｃｉ＋１が一致する候補対しては、データＰ２よりもさらに優先度が低いことを示すデータＰ３が対応づけられる。

【００６７】最後に、「直前の音素」と入力ブロックｃｉ－１とが一致せず、かつ、「直後の音素」と入力ブロックｃｉ＋１とが一致しない候補に対しては、候補として不適当であるため、いかなるデータの対応づけも行わない。

【００６８】上記の分類処理の終了後、音素データ特定部９は、ステップＳ８７において、当該候補が最後の候補か否かを判定する。ここで、最後の候補でない（ＮＯ）と判定されると、ステップＳ８９に制御を移行させ、次の候補の「直前の音素」と「直後の音素」とを読み込み、その後、ステップＳ８５へと制御を移行させる。これ以降、音素データ特定部９は、ステップＳ８５とステップ８７の処理を、ステップＳ８７にて、最後の候補である（ＹＥＳ）と判定されるまで繰返し行う。

【００６９】ステップＳ８７において最後の候補であると判定されると、次のステップＳ９１においては、上記優先度について分類された候補を次のような処理によって絞り込む。

【００７０】音素データ特定部９は、先ず、ステップＳ９１にて、データＰ１が対応づけられた候補があるか否かを判定し、データＰ１が対応づけられた候補がある（ＹＥＳ）場合は、ステップＳ９３において、その他の候補に対して候補除外処理を行う。一方、ステップＳ９

１において、データＰ１が対応づけられた候補がない（ＮＯ）と判定された場合は、次のステップＳ９５において、データＰ２が対応づけられた候補があるか否かを判定する。データＰ２が対応づけられた候補がある（ＹＥＳ）と判定された場合は、ステップＳ９３において、その他の候補に対して候補除外処理を行う。一方、ステップＳ９５にて、データＰ２が対応づけられた候補がない（ＮＯ）と判定された場合は、次のステップＳ９７にて、データＰ３が対応づけられた候補があるか否かを判定する。ここで、データＰ３が対応づけられた候補がある（ＹＥＳ）と判定された場合は、ステップＳ９３において、その他の候補に対して候補除外処理を行う。そして、ステップＳ９７において、データＰ３が対応づけられた候補がない（ＮＯ）と判定された場合、即ち、全ての候補に対し、データＰ１、Ｐ２およびＰ３のいずれのデータの対応づけもなされていない場合は、音素データ特定部９は、ステップＳ９９において、全ての候補について候補除外処理を行う。そして、ステップＳ９９における処理、あるいは、ステップＳ９３における処理の終了により、当該入力ブロックｃｉについての処理を終える。

【００７１】図６に示すフローチャートは、最後の入力ブロック、即ち入力ブロックｃｎに対して、音素データ特定部９が実行する処理を説明するための図であり、本図におけるステップＳ１１１においては、入力ブロックｃｎ－１を主記憶部に読み込む。そして、現在残っている候補のなかの最初の候補（音素データ）の「直前の音素」を読み込む（ステップＳ１１３）。そして、ステップＳ１１１にて読み込んだ入力ブロックｃｎ－１とステップＳ１１３にて読み込んだ「直前の音素」とが一致するか否かをステップＳ１１５にて判定し、ここで一致しない（ＮＯ）と判定された場合は、音素データ特定部９は、ステップＳ１１７において当該候補について候補除外処理を行う。

【００７２】一方、ステップＳ１１５にて、一致する（ＹＥＳ）と判定された場合は、その候補は、候補としての適格性があるので、候補除外処理は行われない。ステップＳ１１５にてＹＥＳと判定された後、あるいは、ステップＳ１１７における候補除外処理の終了後は、音素データ特定部９は、ステップＳ１１９において、当該候補が最後の候補か否かを判定する。ここで、最後の候補でない（ＮＯ）と判定されると、ステップＳ１２１に移行して、次の候補の「直前の音素」を読み込み、そして、音素データ特定部９は、ステップＳ１１５からステップＳ１１９までの一連の処理を、ステップＳ１１９において、最後の候補である（ＹＥＳ）であると判定されるまで順次実行する。

【００７３】従って、図４ないし図６を参照しながら説明したように、この実施の形態の音声合成装置では、音素データ特定部９は、互いに隣接する入力ブロックを、音素データの「直前の音素」あるいは「直後の音素」に対比させながら、候補を選定するため、隣接音声環境が適切な音声素片を抽出することができ、音声素片の接続部分のノイズを低減して、録音された実音声の特徴を備え、連続性が良好な自然な合成音を得られるという効果がある。

【００７４】次に、音素データ特定部９は、上記のように選定された候補を最終的に１つの音素データにまで特定する処理を行う。

【００７５】図７に示すフローチャートにおいて、音素データ特定部９は、先ずステップＳ１３１で、現在残っている候補の数を算出する。次のステップＳ１３３では、音素データ特定部９は、ステップＳ１３１での候補数算出の結果によって、候補が存在するか否かを判定する。このステップにおいて、候補が存在しない（ＮＯ）と判定された場合、即ち、候補がすべて候補除外処理により除外されていた場合は、入力ブロック、平均基本周波数および再生時間長による候補の選定処理（図２のステップＳ９）が終了した時点で候補であったものを、再び候補とする処理を行う（ステップＳ１３５）。

【００７６】一方、ステップＳ１３３にて、候補が存在すると判定された場合（ＹＥＳ）あるいは、ステップＳ１３５での処理の終了後は、音素データ特定部９は、その候補のなかで、入力ブロックに対して割り当てられた平均基本周波数ｆと当該候補の「平均基本周波数」との偏差Δｆが最小となる音素データを最終的に特定する（ステップＳ１３７）。

【００７７】以上のような処理過程を経て、図２に示すステップＳ１１では、各入力ブロックに対して、最終的に１つの音素データが特定される。

【００７８】そして、図２において、ステップＳ１３では、現在処理を行っている入力ブロックが、最後の入力ブロック（第ｎ入力ブロック）であるか否かが判定され、ここで最後の入力ブロックでない（ＮＯ）と判定されると、ステップＳ１５において、次の入力データを主記憶部に読み込む。そして、制御をステップＳ９に移行させ、入力ブロックに対しての候補選定を行う。これ以降は、ステップＳ１３において、最後の入力ブロックである（ＹＥＳ）と判定されるまで、上記処理を順次実行する。

【００７９】そして、ステップＳ１３にて、最後の入力ブロックであると判定されると、次に音声素片抽出部１１により、音声素片の抽出が行われる（ステップＳ１７）。

【００８０】次に、音声素片の抽出過程を図８に示すフローチャートを参照して説明する。このフローチャートは、１つの入力ブロックについての音声素片の抽出過程を示すものである。

【００８１】本図におけるステップＳ１４１において、入力ブロックに対応した特定音素データの「所属音声素

片データ名」、「音素再生開始時間」および「音素再生終了時間」を読み込む。次に、ステップＳ１４３においては、図１に示すように、実音声波形記憶部１７に記憶されたデータベースのなかの「所属音声素片データ名」（ｏｈａｙｏｕｇｏｚａｉｍａｓｕ００１）が示す波形データの範囲において、特定音素データの「音素再生開始時間」の示す時点（３０００）から「音素再生終了時間」の示す時点（４０００）までの部分的な波形データ（音声素片）を所定の作業領域に読み込み、このデータを合成音生成のための所定の保存領域に保存する。そして、音声素片抽出部１１は、以上の処理をすべての入力ブロックについて行う。尚、各入力ブロックに対して抽出され保存された音声素片を、音声素片ＳＵ１，ＳＵ２，…，ＳＵｎとする。

【００８２】次に無音時間計算部１３は、図２に示すように、全ての音声素片の抽出が終了すると、ステップＳ１９おいて、音声素片を接続する際の接続部分に設ける無音時間を計算する。即ち、この実施の形態は、この無音時間の設定により、音声上の人の息継ぎ等の特徴を再現させようとするものである。

【００８３】次に、この計算方法を図９に示すフローチャートを参照して説明する。このフローチャートは、１つの音声素片と、これと隣接する音声素片との間の接続部分に設けられる無音時間の計算方法を示すものである。

【００８４】無音時間計算部１３は、ステップＳ１５１において、最後の入力ブロックを除いて、入力ブロックに対応する特定音素データの「直後の無音部分長」を読み込む（これをデータＡとする）。次に、ステップＳ１５３において、当該入力ブロックｃｉの直後の入力ブロックｃｉ＋１に対応する特定音素データの「直前の無音部分長」を読み込む（これをデータＢとする）。

【００８５】次に、ステップＳ１５５では、条件として、データＡ＝０であり、かつ、データＢ＝０であるか否かが判定される。この条件を満たす（ＹＥＳ）ときは、無音時間計算部１３は、音声素片ＳＵｉと音声素片ＳＵｉ＋１の接続部分の無音時間は０（ｍｓ）とする（ステップＳ１５９）。一方、ステップＳ１５５の条件を満たさない（ＮＯ）とき、即ち、データＡおよびデータＢのいずれか一方が０でないときは、ステップＳ１５７において、無音時間＝（Ａ＋Ｂ）／２（ｍｓ）に設定する。そして、これら無音時間をデータ（無音時間データ）として所定の保存領域に保存する。

【００８６】例えば、実音声「おはようございます（ｏｈａｙｏｕｇｏｚａｉｍａｓｕ）」に対する波形データが実音声波形記憶部１７に記憶され、音素記号／ｕ／に対する「直後の無音部分長」が１５０ｍｓである場合にあって、入力テキスト「おはよう」に続いて入力された「いいてんきですね」の「う」に対して、前記の音素記号／ｕ／が特定されると、前述の無音時間１５０ｍｓについては、続く入力テキスト「いいてんきですね」に対しては短すぎてしまうが、入力テキスト「い」に対する特定音素データの「直前の無音部分長」が例えば５００ｍｓとすれば、上記の無音時間の設定方法により計算される平均、（１５０ｍｓ＋５００ｍｓ）／２＝３２５ｍｓを設定することで、上述の不都合を回避することができる。尚、この実施の形態では、平均を計算する方法を採用したが、例えば、前の音声素片に対する「直後の無音部分長」の方を重視する等の、状況に応じた重み付け演算を行ってもよい。

【００８７】以上の説明から明らかなように、この実施の形態では、音声上の人の息継ぎ等の特徴が再現可能となる効果が得られる。加えて、無音時間を介して接続される、各音声素片を抽出するための各音素データの「直前の無音部分長」、「直後の無音部分長」に重み付けを行うことにより、偏った無音時間が設定されるのを回避して、実音声上では前後しない音声素片どうしを接続する際にあっても、自然な合成音が得られる。

【００８８】そして、無音時間計算部１３は、最後の音声素片ＳＵｎを除く全ての音声素片に対して上記処理を行うことで、音声素片の接続部分に設ける無音時間を決定し、これらデータを主記憶部の所定保存領域に記憶する。

【００８９】最後に、音声合成部１５は、図２のステップＳ２１において、主記憶部の所定領域に保存された各音声素片と、接続部分に設ける無音時間を示す無音時間データとを読み出し、これらを順次接続して１つの合成音の波形データとする。そして、ステップＳ２３において、合成された波形データをＡ／Ｄ変換してアナログ信号に変換し、そのアナログ信号を増幅して合成音として出力する。

【００９０】従って、本発明に係る音声合成装置の第１の実施の形態によれば、音声波形が予め記憶された実音声波形記憶部１７が設けられ、さらに、音素データ記憶部１９には、実音声波形記憶部１７の音声波形に含まれる音声素片（第１の音声素片）に対して付与され「保存領域１」に保存された音素記号（第１の音素記号）と、その第１の音声素片に隣接する音声素片に対して付与され「保存領域５」あるいは「保存領域６」に保存された「直前の音素」あるいは「直後の音素」、即ち第２の音素記号と、「保存領域２」の「所属音声素片データ名」、「保存領域３」の「音素再生開始時間」並びに「保存領域４」の「音素再生終了時間」からなる第１の音声素片の記憶場所情報とを含む音素データが予め記憶され、音声合成指令として入力音素記号列（入力ブロック）が入力されると、音素データ特定部９により、その入力音素記号列とこれに隣接する入力音素記号列とからなる組と、音素データ記憶部１９における第１の音素記号および第２の音素記号の組とが照合され、入力音素記号列に対する音素データが特定され、その後、音声素片

17

抽出部１１により、実音声波形記憶部１７のなかの特定された音素データの示す記憶場所から、逐次音声素片が抽出され、音声合成部１５により、所定の順序でこれら音声素片が接続され、合成音が出力されるため、音声素片の接続部分のノイズを低減するとともに、録音された実音声の特徴を備えた高品質の合成音出力が可能となる。

【００９１】次に、本発明に係る音声合成装置の第２の実施の形態について説明する。

【００９２】この実施の形態は、上記第１の実施の形態と同様に高品質の合成音を得るとともに、合成音を構成する音声素片の出力タイミングに係る情報を含む文字列Ｌ（以下、リップシンク信号列Ｌという）を生成することを特徴とする。即ち、この実施の形態では、このリップシンク信号列Ｌを、例えばＣＧを制作するための画像処理装置などで解読して、顔画像における口の形状を変化させるタイミングを決定することができるようにしている。

【００９３】具体的には、第２の実施の形態の音声合成装置には、タイミング情報生成部２１が設けられ、このタイミング情報生成部２１においてリップシンク信号列Ｌが生成される。

【００９４】図１０は、このリップシンク信号列Ｌを生成するための処理を示すフローチャートであり、このフローチャートを参照しながら、入力テキスト「こんにちは」に係る処理を説明していく。

【００９５】先ず、上記第１の実施の形態と同様に、入力テキスト「こんにちは」は、図１に示す音素分解処理部３によって、４個の入力ブロック／ｋｏ／，／ｎ／，／ｎｉ／，／ｔｉｗａ／に変換される。そして、音素データ候補選定部７および音素データ特定部９によって、音素データ記憶部１９の中から、各入力ブロックに対して音素データが特定され、無音時間計算部１３によって、入力ブロック／ｋｏ／と／ｎ／のそれぞれに対応する音声素片の接続部分の無音時間５００ｍｓが決定される。

【００９６】タイミング情報生成部２１は、図１０のステップＳ１６１において、各入力ブロックに対して特定された音素データの「音素再生開始時間」と「音素再生終了時間」との差を再生時間長として演算し、入力ブロックと再生時間長とからなる組を再生される順序でメモリなどに保存することにより、表１に示すような「合成音再生データ」を構築する。このとき、タイミング情報生成部２１は、無音時間５００ｍｓを１つの音声素片の再生時間長とみなし、これに「無音」を表す／ｑ／を入力ブロックとして対応づける。

【００９７】

【表１】

18

合成音再生データ

i	1	2	3	4	5
ci	/ko/	/q/	/n/	/ni/	/tiwa/
ti	2500	500	1000	2000	4000

表１においてｉは、再生の順序を表す変数であり、この実施の形態では、変換された４個の入力ブロックと前述の／ｑ／とを合わせて、その最大値ｎ＝５となっている。また、ｃｉは各入力ブロックを示す変数であり、そして、ｔｉは、各入力ブロックに対して抽出された音声素片の再生時間長を示す変数であって、ｔ２としての無音時間５００ｍｓが含まれている。

【００９８】図１０に戻り、タイミング情報生成部２１は、ステップＳ１６３において、リップシンク信号列Ｌの内容をクリアして、何も文字が含まれていない状態（図中では、記号”　”で示す）とする。また、先頭の音声素片の出力開始時からの累積経過時間を表す変数Ｔを０（零）にリセットするとともに再生順序を表す変数ｉを１とする。即ち、これらの処理が初期化動作として実行される。

【００９９】そして、タイミング情報生成部２１は、ステップＳ１６５において、現在のリップシンク信号列Ｌに対して、入力ブロックｃｉを右側から結合する（以下、単に「右結合する」という）。尚、このステップＳ１６５の処理が行われる前においては、Ｌ＝”　”であるから、このステップＳ１６５の処理後において、Ｌ＝”ｋｏ”となる。尚、結合時には、入力ブロックの前後の記号／は取除かれる。

【０１００】続くステップＳ１６７では、ｉ＝１であるか否かが判定される。ここで、ｉ＝１（ＹＥＳ）と判定された場合は、ステップＳ１６９にて、タイミング情報生成部２１は、リップシンク信号列Ｌに文字”０（零）”を右結合する。

【０１０１】その後ステップＳ１７５へと進み、ここで、処理中の当該入力ブロックｃｉが最後の入力ブロックである（変数ｉ＝ｎ）か否かが判定され、ここでＮＯと判定されたときは、ステップＳ１７７にて、変数ｉが１繰上げられ（ｉ＝ｉ＋１）、再びステップＳ１６５の処理が行われる。

【０１０２】変数ｉが繰上げられてｉ＝２となったときは、ステップＳ１６５にて入力ブロック／ｑ／が右結合され、リップシンク信号列Ｌ＝”ｋｏ０ｑ”となる。続くステップＳ１６７では、ｉ≠１（ＮＯ）と判定されるため、ステップＳ１７１へと進む。

【０１０３】このステップＳ１７１では、タイミング情報生成部２１は、累積経過時間を表す変数Ｔに対して、現在の変数ｉの１つ前のｉ－１に対応する再生時間長ｔ（ｉ－１）を加算処理する。即ち、ｔ１＝２５００が加算され、変数Ｔ＝２５００となる。続いて、ステップＳ１７３では、タイミング情報生成部２１は、変数Ｔの内

容を文字列に変換する処理を行い、変換後の文字列ｃｈ（Ｔ）をリップシンク信号列Ｌに対して右結合する。従って、ｉ＝２のときは、ステップＳ１７３の処理後のリップシンク信号列ＬはＬ＝”ｋｏ０ｑ２５００”となる。

【０１０４】このようにして、タイミング情報生成部２１は、リップシンク信号列Ｌを”ｋｏ０ｑ２５００”→”ｋｏ０ｑ２５００ｎ３０００”→”ｋｏ０ｑ２５００ｎ３０００ｎｉ４０００”と生成していき、最終的にリップシンク信号列Ｌ＝”ｋｏ０ｑ２５００ｎ３０００ｎｉ４０００ｔｉｗａ６０００”を完成させる。そして、ステップＳ１７５において当該入力ブロックｃｉが最後の入力ブロックである（ｉ＝ｎ）と判定されると、このリップシンク信号列Ｌの生成処理が終了する。

【０１０５】こうして生成されたリップシンク信号列Ｌに含まれる各数値は、合成音における先頭の音声素片の出力開始時間（０）と、これを基準とした２番目以降の各音声素片の出力開始時間を表している。即ち、本例では、入力ブロックｃ１＝／ｋｏ／に対応する音声素片「こ」の出力開始時間を基準時とすると、２５００ｍｓ後に無音状態となり、基準時から３０００ｍｓ後に音声素片「ん」が出力され、基準時から４０００ｍｓ後に音声素片「に」が出力され、最後に音声素片「ちわ」が、基準時から６０００ｍｓ後に出力されることを示している。

【０１０６】その後、生成されたリップシンク信号列Ｌは、画像処理装置などにて解読され、顔画像における口の形状を変化させるタイミングが決定される。

【０１０７】従って、本発明に係る音声合成装置の第２の実施の形態によれば、タイミング情報生成部２１を設けたことにより、音声素片の出力タイミングに係る情報を含むリップシンク信号列Ｌを生成することができ、このリップシンク信号列Ｌを利用して、音声波形を解析することなく、顔画像における口の形状を最適なタイミングで変化させることができる。

【０１０８】尚、本例の音声素片「ちわ」のように、音声素片が複数の音素からなる場合にあっては、その音声素片に、最初の音素、本例でいえば、／ｔ／に対する口の形状を対応させるか、あるいは最後の音素、本例でいえば、／ａ／に対する口の形状を対応させるかの２通りの方法があるが、後者を選択することにより違和感の少ない顔の表情を得ることができる。

【０１０９】また、上記のリップシンク信号列Ｌにおける文字の並び方は、音声素片が出力されるタイミングを判読しやすいものとしたが、必ずしもこの並び方に限定されるものではなく、所定の規則により、音声素片が出力されるタイミングを解読可能とさえすればよい。

【０１１０】ところで、上記第２の実施の形態では、リップシンク信号列Ｌに含まれるタイミングに係る情報を、各音声素片ごとに求めたが、音声素片よりも短い音

声ごとの再生時間長を所定の規則により求め、この再生時間長を用いて、音声素片よりも短い音声との同期をとるためのタイミング情報をリップシンク信号列Ｌに含ませることにより、このリップシンク信号列Ｌを用いて、画像処理装置などにおいて、顔画像における口の形状をより滑らかに変化させることができる。

【０１１１】そこで、音声素片よりも短い音声との同期をとるための情報を含むリップシンク信号列Ｌを生成可能とした本発明の音声合成装置の第３の実施の形態を説明する。

【０１１２】図１１は、第３の実施の形態の構成の一部を示すブロック図である。

【０１１３】この実施の形態では、リップシンク信号生成部２１には、再生時間演算部２１ａが設けられ、ここには、予め、母音および子音それぞれの再生時間長を求めるための比率ｘおよびｙが設定されている。

【０１１４】次に、第３の実施の形態における音声合成装置の動作を説明する。

【０１１５】タイミング情報生成部２１は、先ず、前記の表１に示すような合成音再生データをメモリなどから読込み、この合成音再生データに含まれる入力ブロックの中から、適宜、分解が必要な入力ブロックを選択して、音素分解処理部３へと供給する。

【０１１６】具体的には、例えば、再生時間長ｔｉが長い場合は、音声と画像の同期が不自然に感じられやすいため、所定の再生時間長以上のデータを選択して、音素分解処理部３へと供給すればよい。前述の合成音「こんにちは」を例にすると、音声素片「ちわ」の再生時間長ｔ５が比較的長いため、入力ブロック／ｔｉｗａ／が音素分解処理部３へと供給される。

【０１１７】音素分解処理部３は、供給された「分解前の入力ブロック」を音素記号に分解する。この過程において、音素分解処理部３は、「分解前の入力ブロック」から、母音と子音とからなる音節に相当する音節記号を抽出し、さらに、音節記号を母音あるいは子音に相当する音素記号に分解する。

【０１１８】そして、音素分解処理部３は、この分解処理によって得られた音素記号に、音節の区切を解析できるような音節区切データを付加して、タイミング情報生成部２１に返送する。この音節区切データは、例えば、”１２，３４”のようなデータに、「１番目と２番目の音素記号が、母音および子音にそれぞれ相当する１つの音節記号を構成し、これに対して，（カンマ）で区切られた３番目と４番目の音素記号がもう１つの音節記号を構成する」という意味をもたせることによって生成することが可能である。

【０１１９】尚、音節とは、必ずしも母音と子音との組合わせに限られるものではなく、例えば、音声「ん」に相当する音素記号「ｎ」を音節とすることもできる。本例では、「分解前の入力ブロック」ｃ５＝／ｔｉｗａ／

21

が、／ｔ／，／ｉ／，／ｗ／，／ａ／の各音素記号に分解され、前述の音節区切データ”１２，３４”とともにタイミング情報生成部２１へと返送される。

【０１２０】続いて、タイミング情報生成部２１の再生時間演算部２１ａは、音節区切データを解析し、音素分解処理部３から返送された音素記号に複数の音節記号が含まれるような場合にあっては、その各音節記号に対し当該各音節記号に対応する音節の再生時間長を与える。本例では、分解前の入力ブロックｃ５＝／ｔｉｗａ／には、／ｔｉ／および／ｗａ／の２個の音節記号が含まれるので、これら各音節記号に対して、均等な音節の再生時間長が与えられる。即ち、音節の再生時間長＝（分解前のｔ５）／（音節数）＝４０００／２＝２０００（ｍｓ）となる。このようにして、本実施の形態では、音節との同期をとるためのタイミング情報をリップシンク信号列Ｌに含ませることができるようになっている。

【０１２１】そして、再生時間演算部２１ａは、各音節記号の中の音素記号／ｔ／，／ｉ／，／ｗ／，／ａ／のうちの母音記号に対しては、音節の再生時間長に対する母音の再生時間長の比率として設定された比率ｘを与える一方、子音記号に対しては、比率ｙ（ただしｙ＝１－ｘ）を与え、これら比率によって、音声素片に含まれる＊

22

＊母音／子音の再生時間長を演算することで、音素との同期をとるためのタイミング情報をリップシンク信号列Ｌに含ませることができるようになっている。

【０１２２】具体的には、ｘ＝０．４とし、ｙ＝０．６とした場合に、母音および子音の再生時間長は以下のように演算される。

【０１２３】母音の再生時間長＝音節の再生時間×比率ｘ＝２０００×０．４＝８００（ｍｓ）

子音の再生時間長＝音節の再生時間×比率ｙ＝２０００×０．６＝１２００（ｍｓ）

タイミング情報生成部２１は、演算された各再生時間長と各音素記号とを互いに対応づけて、分解前の再生時間長および入力ブロックに置換えることで、合成音再生データを再構築する。

【０１２４】具体的には、表２に示すような合成音再生データが再構築される。即ち、表１に示した合成音再生データにおける５番目の再生時間長および入力ブロックが、表２に示す５番目ないし８番目の再生時間長およびその入力ブロックに置換えられる。

【０１２５】

【表２】

合成音再生データ

i	1	2	3	4	5	6	7	8
ci	/ko/	/q/	/n/	/ni/	/t/	/i/	/w/	/a/
ti	2500	500	1000	2000	800	1200	800	1200

合成音再生データの再構築後は、上記第２の実施の形態と同様に、タイミング情報生成部２１によって、合成音再生データからリップシンク信号列Ｌが生成される。表２の再構築後の合成音再生データからは、リップシンク信号列Ｌ＝”ｋｏ０ｑ２５００ｎ３０００ｎｉ４０００ｔ６０００ｉ６８００ｗ８０００ａ８８００”が生成される。そして、このリップシンク信号列Ｌが画像処理装置などにて解読され、顔画像における口の形状を変化させるための、より詳細なタイミングが決定される。尚、「分解前の入力ブロック」を音素記号まで分解せずに、音節記号までの分解にとどめることも勿論可能である。

【０１２６】以上説明したように、この第３の実施の形態によれば、合成音として出力される音声素片に、母音あるいは子音などの音素、あるいは母音および子音からなる音節などの、音声素片よりも短い音声が含まれる場合にあっては、再生時間演算部２１ａにより、その音声ごとの再生時間長が演算され、タイミング情報生成部２１によって、音声素片よりも短い音声との同期をとるための情報を含むリップシンク信号列Ｌが生成されるため、画像処理装置などにおいて、顔画像における口の形状をさらに細かく、滑らかに変化させることができる。

【０１２７】また、音声素片に母音および子音からなる音節が含まれるときにあっては、予め設定された、音節の再生時間長に対する母音および子音それぞれの再生時間長の比率ｘおよびｙによって、母音および子音のそれぞれの再生時間長が演算され、母音あるいは子音と同期をとるための情報がリップシンク信号列Ｌに含まれることとなるため、顔画像における口の形状を母音あるいは子音と同期をとって変化させることができる。また、比率ｘ，ｙの設定を変えることによって、母音あるいは子音と同期をとるための情報を調整することができる。

【０１２８】尚、本発明に係る音声合成装置は、上記実施の形態に限るものではなく、ソフトウェアで構成された上記各処理を実行する部分を、機械読み取り可能な記録媒体に記録することも可能である。また、上記実施の形態は、音声合成の方法あるいはタイミングに係る情報の生成方法などのアルゴリズムとしても優れたものである。

【０１２９】そして、上記各機能を集積回路等のハードウェアで構成しても同様の効果を得ることができ、さらに、単一のコンピュータ上でのみならず、ネットワークを構成する端末やサーバマシン、あるいは、画像処理装置等に各機能を分散配備させてもよい。

【０１３０】

【発明の効果】以上説明したように、本発明の請求項１に係る音声合成装置によれば、音素データ特定手段により、入力音素記号列と隣接入力音素記号列とからなる組と、音素データに含まれる第１のラベルおよび第２のラ

ベルの組とが照合され、入力音素記号列に対して、隣接する音声環境の最適な音声データが特定されるため、音声素片の接続部分のノイズが低減されるとともに、実音声の特徴を備えた高品質の合成音出力が可能となる。

【０１３１】また、本発明の請求項２に係る音声合成装置によれば、音声素片の接続部分の状態が、音素データに付与された時間情報に基づいて設定されるため、実音声上の息継ぎ等の特徴が再現可能となる。

【０１３２】また、本発明の請求項３に係る音声合成装置によれば、無音部分の時間情報に対して重み付けが行われ、音声素片を接続する部分に設ける無音時間が算出されるため、どちらか一方の音声素片の特徴に偏った無音時間が設定されることが回避され、たとえ実音声上で隣接していない音声素片どうしを接続する際にあっても、自然な合成音が得られる。

【０１３３】また、本発明の請求項４に係る音声合成装置によれば、実音声上では連続することが極めて希な同一の音声素片の特定が規制されて、合成音が不自然になる不都合が回避される。

【０１３４】また、本発明の請求項５に係る音声合成装置によれば、タイミング情報生成手段により音声素片の出力タイミングに係る情報が生成されるため、音声波形を解析することなく、顔画像の口の形状を合成音に同期させて変化させることができる。

【０１３５】また、本発明の請求項６に係る音声合成装置によれば、タイミング情報生成手段により、音声素片よりも短い音声との同期をとるための情報が出力タイミングに係る情報に含ませられるため、口の形状をより滑らかに変化させることができる。

【図面の簡単な説明】

【図１】本発明に係る音声合成装置の第１の実施の形態を示す図である。

【図２】図１に示した形態における音声合成に関する処理過程を示したフローチャートである。

【図３】図１に示した形態における入力ブロック、平均基本周波数および再生時間長に基づく音声データの候補選定過程を示したフローチャートである。

【図４】図１に示した形態における第１入力ブロックについての候補選定過程を示したフローチャートである。

【図５】図１に示した形態における第１入力ブロックおよび最後の入力ブロックを除く入力ブロックについての候補選定過程を示したフローチャートである。

【図６】図１に示した形態における最後の入力ブロックについての候補選定過程を示したフローチャートである。

【図７】図１に示した形態における入力ブロックについての音素データの特定過程を示したフローチャートである。

【図８】図１に示した形態における音声素片の抽出過程を示したフローチャートである。

【図９】図１に示した形態における無音時間の計算過程を示したフローチャートである。

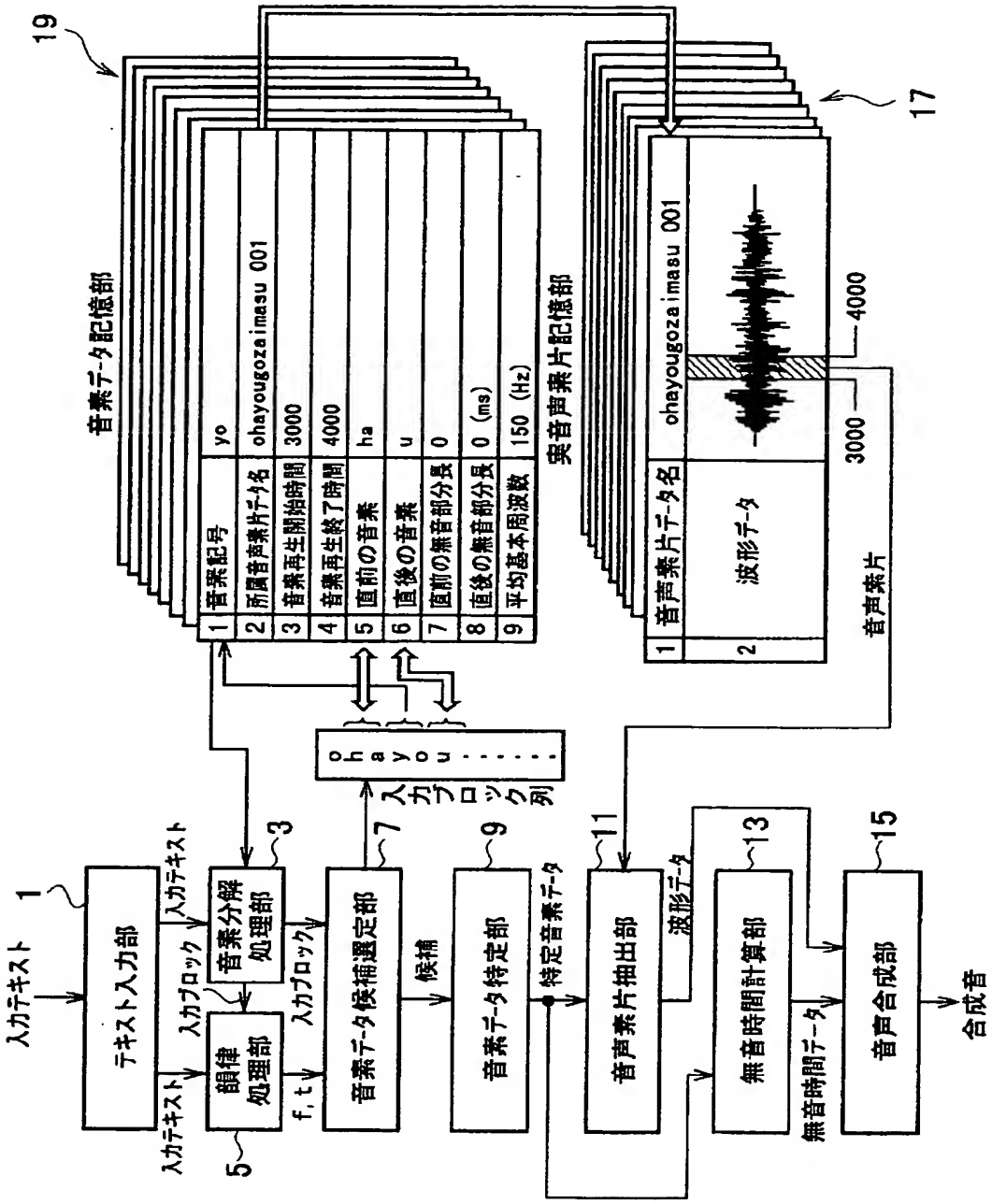
【図１０】本発明に係る音声合成装置の第２の実施の形態の動作を示すフローチャートである。

【図１１】本発明に係る音声合成装置の第３の実施の形態の構成の一部を示すブロック図である。

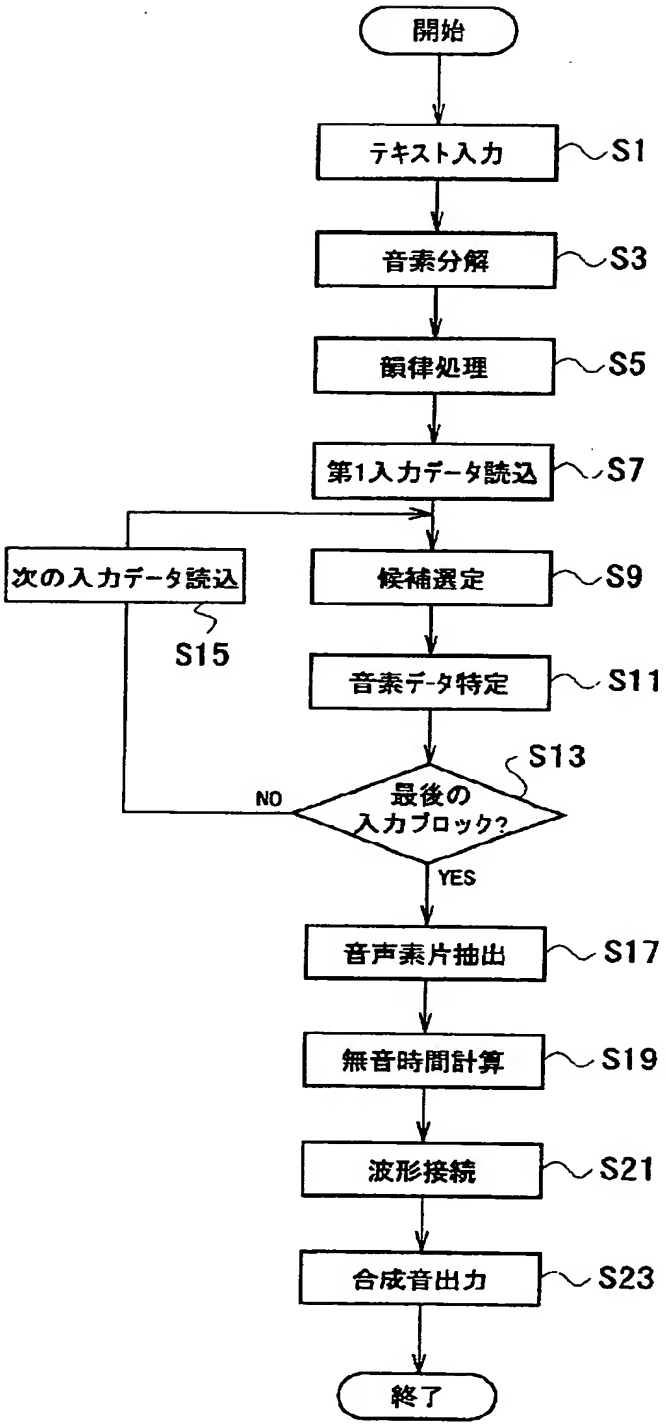
【符号の説明】

１　テキスト入力部
３　音素分解処理部
５　韻律処理部
７　音素データ候補選定部
９　音素データ特定部
１１　音声素片抽出部
１３　無音時間計算部
１５　音声合成部
１７　実音声波形記憶部
１９　音素データ記憶部
２１　タイミング情報生成部
２１ａ　再生時間演算部

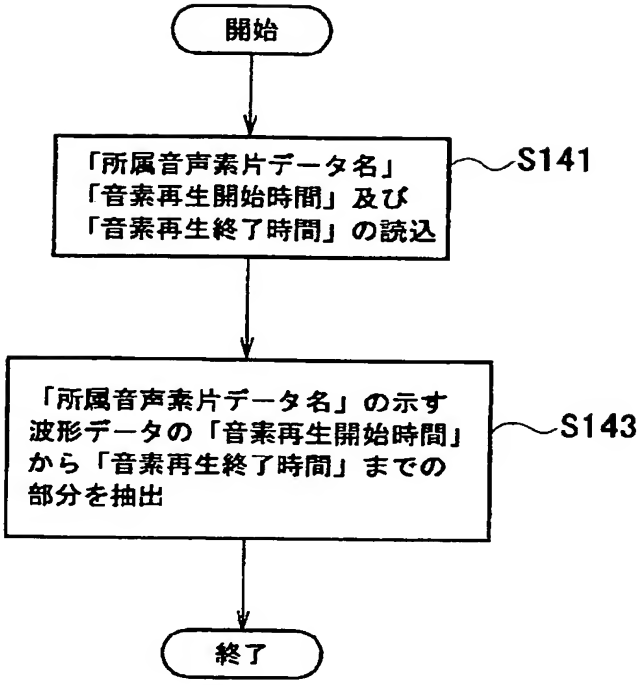
【図１】


19
音素データ記憶部
1 音素記号 yo
2 所属音声素片データ名 ohayougozaimasu 001
3 音素再生開始時間 3000
4 音素再生終了時間 4000
5 直前の音素 ha
6 直後の音素 u
7 直前の無音部分長 0
8 直後の無音部分長 0 (ms)
9 平均基本周波数 150 (Hz)
17
実音声素片記憶部
1 音声素片データ名 ohayougozaimasu 001
2 波形データ
4000
3000
音声素片
o h a y o u ・・・・・
入力ブロック列
入力テキスト 1
テキスト入力部
入力テキスト
入力ブロック
音素分解処理部 3
入力テキスト
韻律処理部 5
f, t
入力ブロック
音素データ候補選定部 7
候補
音素データ特定部 9
特定音素データ
音声素片抽出部 11
波形データ
無音時間計算部 13
無音時間データ
音声合成部 15
合成音

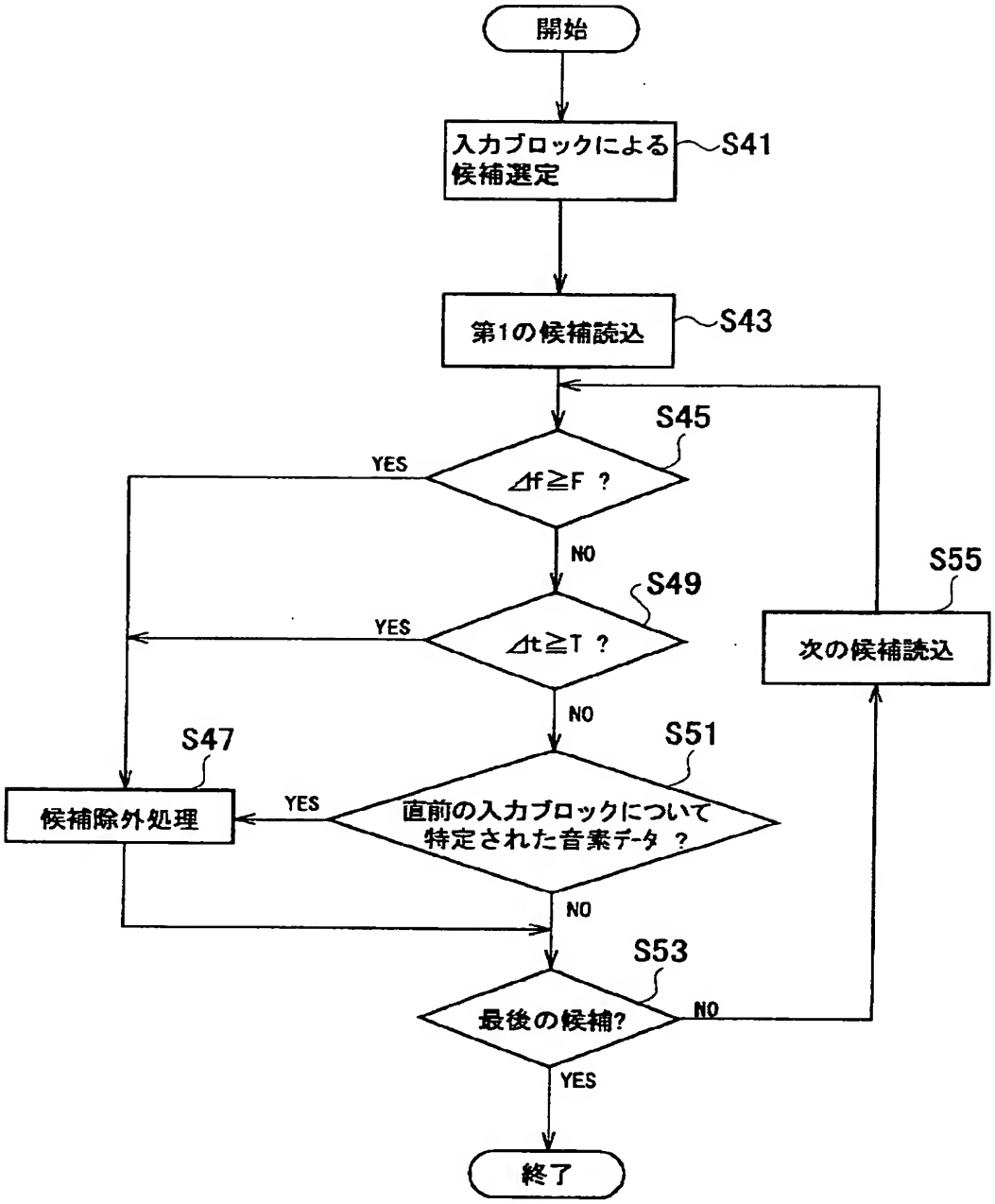
【図２】


開始
テキスト入力
S1
音素分解
S3
韻律処理
S5
第1入力データ読込
S7
次の入力データ読込
S15
候補選定
S9
音素データ特定
S11
S13
NO
最後の
入力ブロック?
YES
音声素片抽出
S17
無音時間計算
S19
波形接続
S21
合成音出力
S23
終了

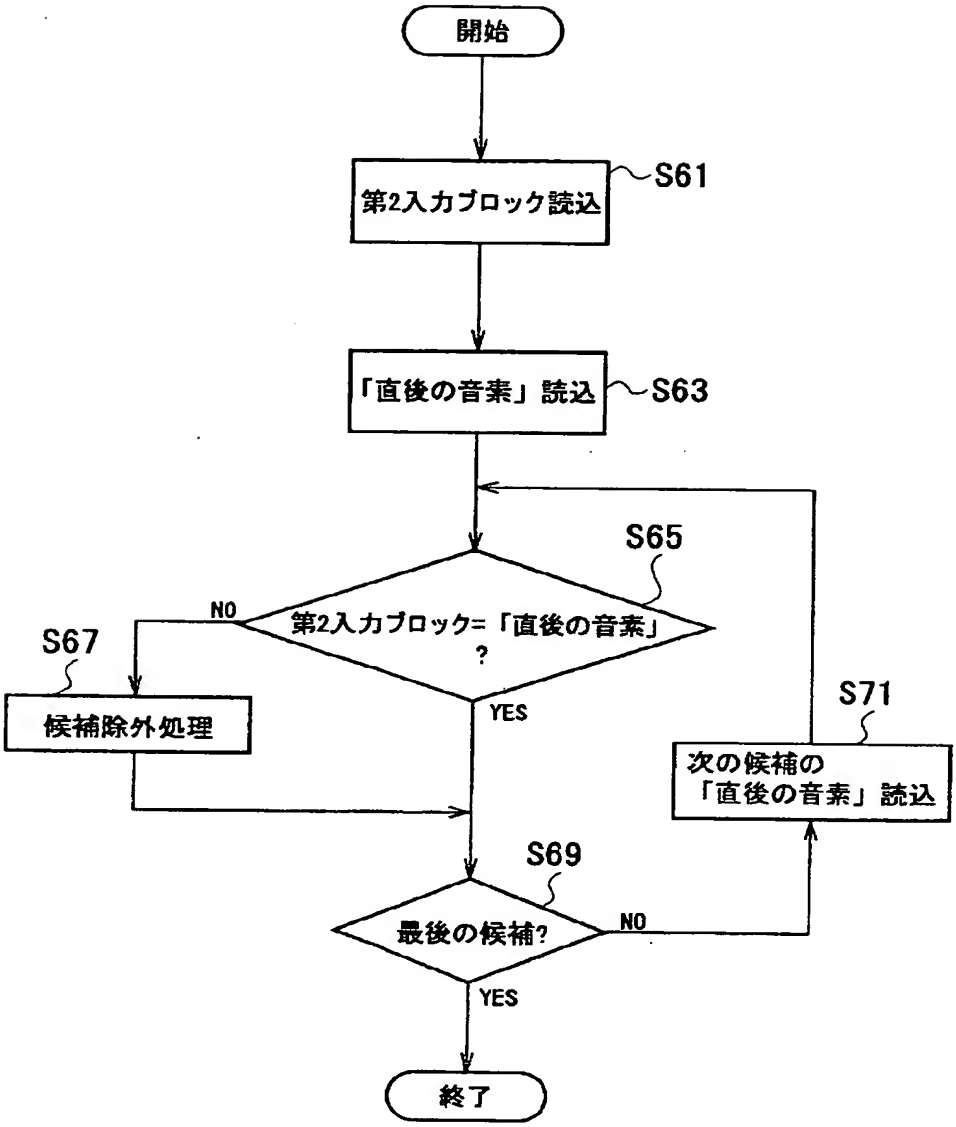
【図８】


開始
「所属音声素片データ名」
「音素再生開始時間」及び
「音素再生終了時間」の読込
S141
「所属音声素片データ名」の示す
波形データの「音素再生開始時間」
から「音素再生終了時間」までの
部分を抽出
S143
終了

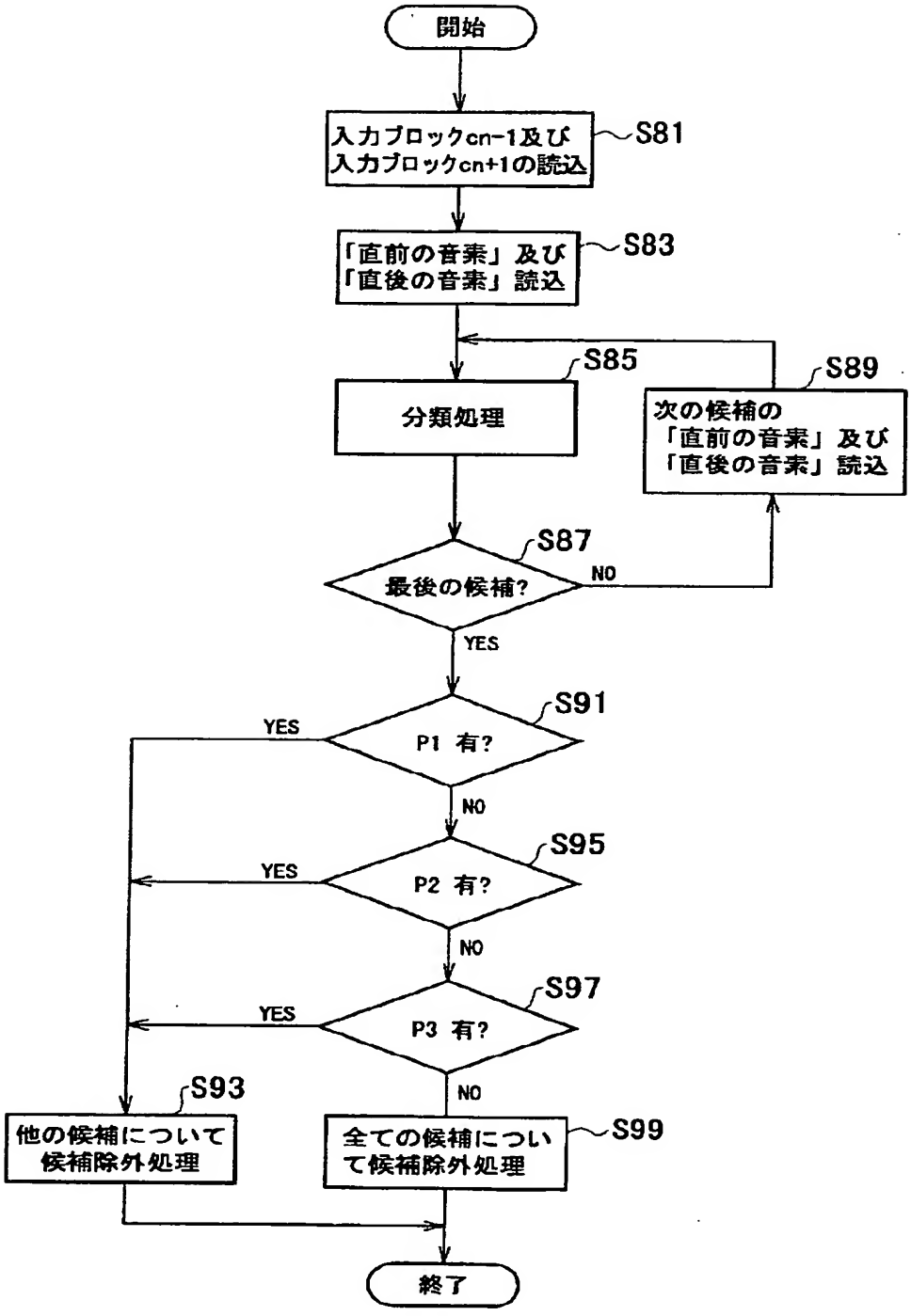
【図３】


開始
入力ブロックによる候補選定
S41
第1の候補読込
S43
S45
⊿f≧F ？
YES
NO
S49
⊿t≧T ？
YES
NO
S51
直前の入力ブロックについて特定された音素データ ？
YES
NO
S47
候補除外処理
S53
最後の候補?
NO
YES
S55
次の候補読込
終了

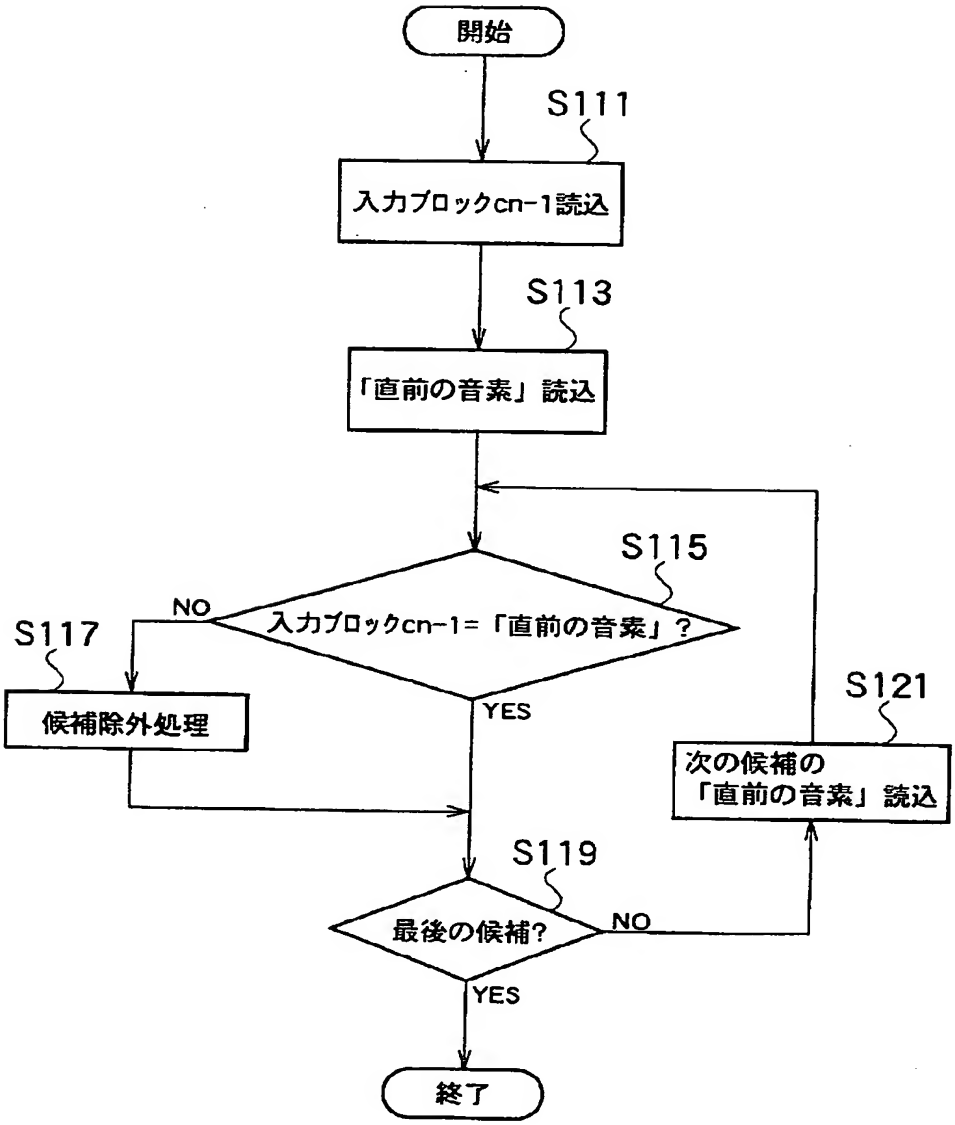
【図４】


開始
第2入力ブロック読込
S61
「直後の音素」読込
S63
S65
第2入力ブロック=「直後の音素」
?
NO
YES
S67
候補除外処理
S71
次の候補の
「直後の音素」読込
S69
最後の候補?
NO
YES
終了

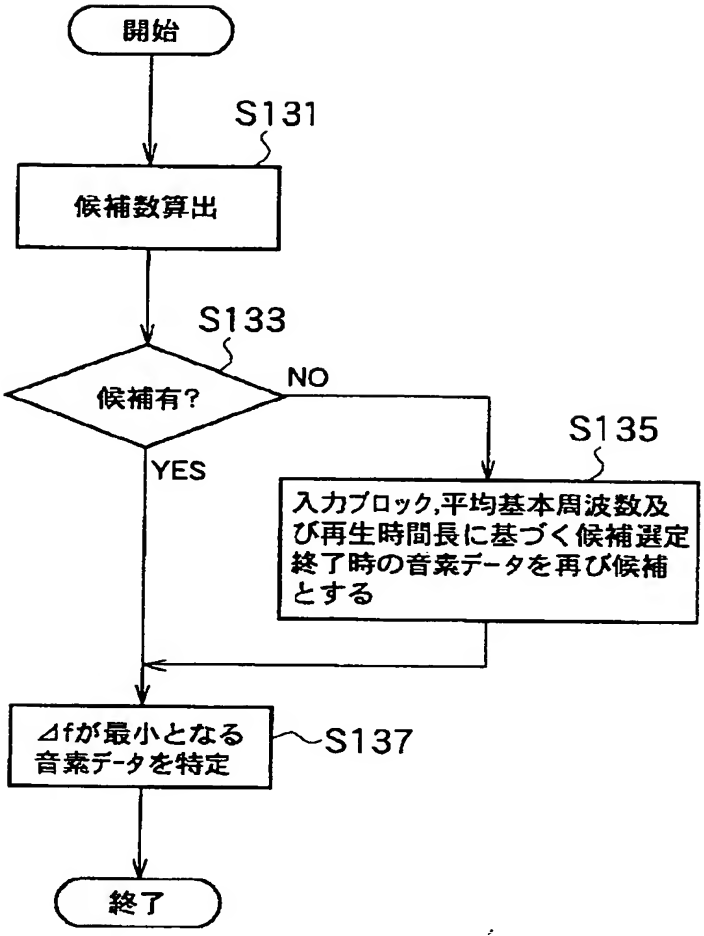
【図５】


開始
入力ブロックcn-1及び
入力ブロックcn+1の読込
S81
「直前の音素」及び
「直後の音素」読込
S83
S85
分類処理
S89
次の候補の
「直前の音素」及び
「直後の音素」読込
S87
最後の候補?
NO
YES
S91
P1 有?
YES
NO
S95
P2 有?
YES
NO
S97
P3 有?
YES
NO
S93
他の候補について
候補除外処理
全ての候補につい
て候補除外処理
S99
終了

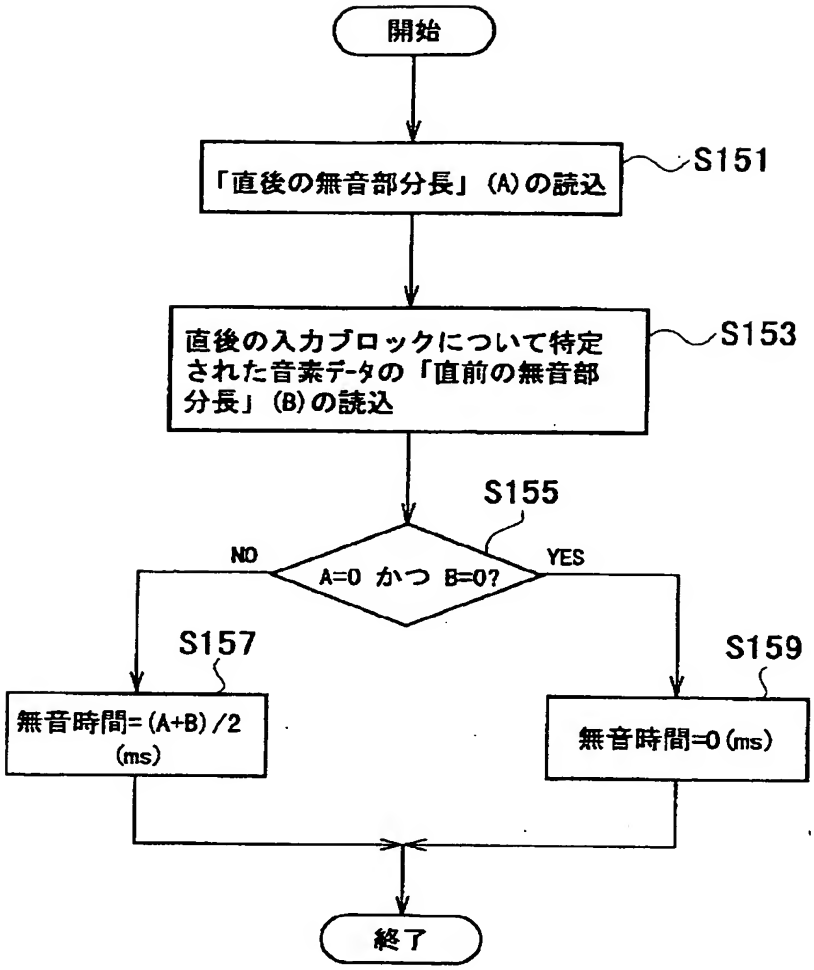
【図６】


開始
S111
入力ブロックcn-1読込
S113
「直前の音素」読込
S115
入力ブロックcn-1=「直前の音素」？
NO
S117
候補除外処理
YES
S121
次の候補の
「直前の音素」読込
S119
最後の候補?
NO
YES
終了

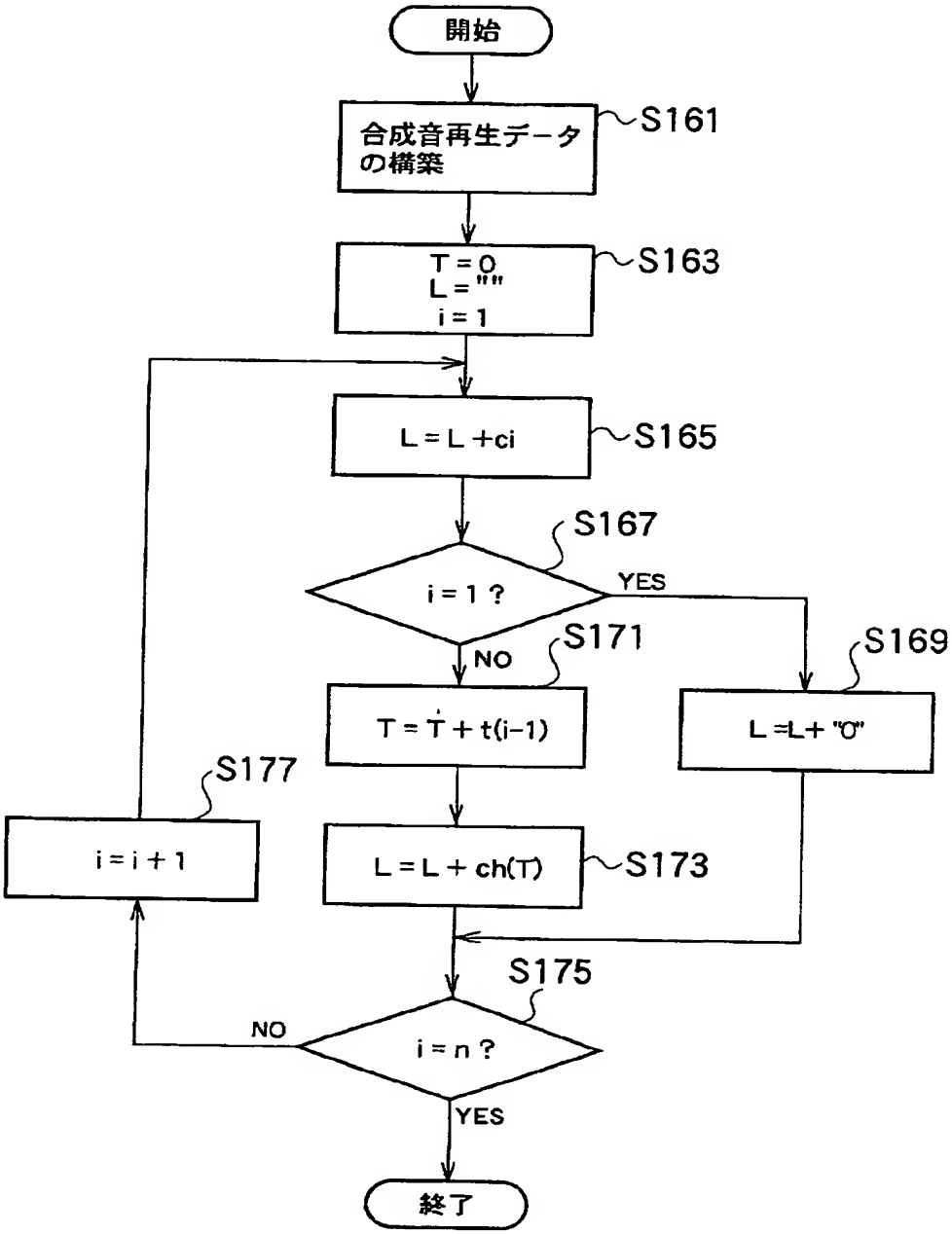
【図７】


開始
S131
候補数算出
S133
候補有?
NO
YES
S135
入力ブロック,平均基本周波数及び再生時間長に基づく候補選定終了時の音素データを再び候補とする
⊿fが最小となる音素データを特定
S137
終了

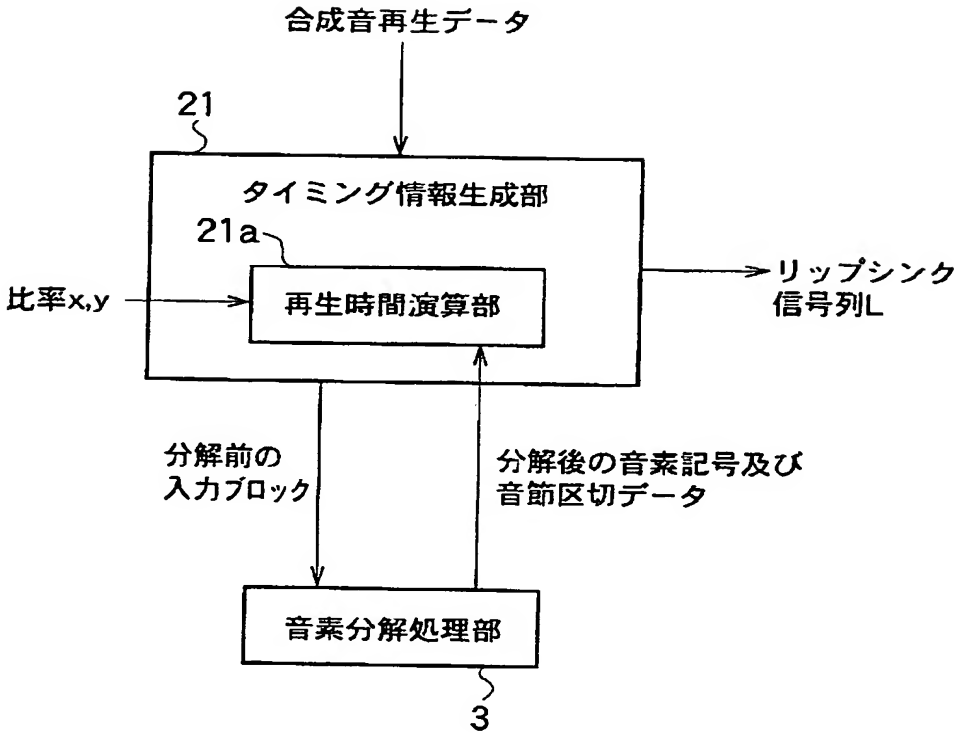
【図９】


開始
「直後の無音部分長」(A)の読込
S151
直後の入力ブロックについて特定された音素データの「直前の無音部分長」(B)の読込
S153
S155
A=0 かつ B=0?
NO
YES
S157
無音時間=(A+B)/2 (ms)
S159
無音時間=0(ms)
終了

【図１０】


開始
合成音再生データの構築
S161
T＝0
L＝""
i＝1
S163
L＝L＋ci
S165
S167
i＝1？
YES
NO
S171
S169
T＝T＋t(i−1)
L＝L＋"0"
S177
i＝i＋1
L＝L＋ch(T)
S173
S175
i＝n？
NO
YES
終了

【図１１】


合成音再生データ
21
タイミング情報生成部
21a
比率x,y
再生時間演算部
リップシンク
信号列L
分解前の
入力ブロック
分解後の音素記号及び
音節区切データ
音素分解処理部
3